

A THREE-STAGE TEXT NORMALIZATION STRATEGY FOR MANDARIN TEXT-TO-SPEECH SYSTEMS

Tao Zhou¹, Yuan Dong^{1,2}, Dezhi Huang², Wu Liu², Haila Wang²

¹Beijing University of Posts and Telecommunication, Beijing
²France Telecom R&D, Beijing

ABSTRACT

Text normalization is an important component in mandarin Text-to-Speech system. This paper develops a taxonomy of Non-Standard Words (NSW's) based on a Large-scale Chinese corpus and proposes a three-stage text normalization strategy: Finite State Automata (FSA) for initial classification, Maximum Entropy (ME) Classifier & Rules for further classification and General Rules for standard word conversion. The three-stage approach achieves Precision of 96.02% in experiments, 5.21% higher than that of simple rule based approach and 2.21% higher than that of simple machine learning method. Experiments results show that the approach of three-stage disambiguation strategy for text normalization makes considerable improvement, and works well in real TTS system.

Index Terms— Text-to-Speech, Text Normalization, Finite State Automata (FSA), Maximum Entropy (ME) Classifier, Standard Word Conversion

1. INTRODUCTION

Text-to-speech is an important technique to generate the artificial speech from text dependent application. This technique has been widely applied in many fields, such as telecommunication services, embedded mobile application and entertainment. Real text contains many Non-Standard Words (NSW's), including numbers, symbols and some alphabets [1]. However, NSW cannot be detected by an application of "letter-to-sound" and may be recognized as different standard words depending on both the local text and the text genre. So it is in general a very hard homograph disambiguation task [2]. In the current text analysis field, text normalization is considered as a crucial component of text analysis in TTS [3]. In Nuance Vocalizer, over 20% of the core application code (line of code metric) is devoted to text normalization, and the new input forms continue to be added [4].

Typical methods for text normalization are based on simple handcrafted rules [5]. Such rules usually have considerable amount of codes, and have difficulty in management and adaptation to new domains.

On the other hand, in view of homograph disambiguation, many machine learning methods are employed and have shown their advantages [6]. Decision tree and decision list are used in

English and Hindi text normalization [7]. Winnow is used for homograph disambiguation in Thai text analysis [8]. Support Vector Machine is applied to Persian NSW's classification [9]. These machine learning methods are better in management, but cannot make a general covering of all NSW's in Chinese text.

A designing of combining machine learning methods and rules can probably be more effective in dealing with NSW's [10]. The approach proposed in this paper utilizes a three-stage text normalization strategy. First of all, Finite State Automata detects NSW's from the real text and makes an initial classification. Then, Maximum Entropy classifier and rules are used for further classification; the maximum entropy classifiers are used for some general types of NSW's while rules make a supplement. Finally, a conversion module based on general rules is applied to transform NSW's to pronouncing text for the following segmentation process in TTS system.

The rest of the paper is organized as follows. Section 2 describes the main proposed approach of normalization module in details. Section 3 shows the actual application in TTS system and the experiment result. Conclusion is given in Section 4.

2. DESCRIPTION OF THE APPROACH

Based on the taxonomy developed by a systematic investigation of a large scale corpus, People Daily Corpus, a three-stage mandarin text normalization process is designed. The entire procedure is shown in Figure 1. Here, SSML means Speech Synthesis Markup Language.

Finite State Automata classifiers are used for NSW's detection and initial classification. Rules and Maximum Entropy classifiers are certainly applied for subclass disambiguation and further classification. Finally, General Rules are used to generate standard words for latter process of TTS system.

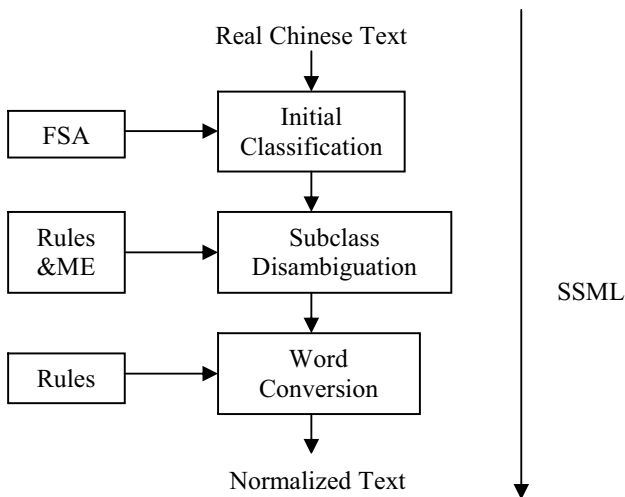


Fig.1.Flow of text normalization

2.1. Finite State Automata Classifiers

FSA Classifiers are of great importance in text normalization. It defines categories of words to process, according to detection, classification and transformation.

Table 1.Format of Not-Standard Word (NSW)

Numbers	digits	110, 911, ...
	colon	12:12, 15:30, ...
	dot	1.23, 12.9, ...
	range	10-20, ...
	other	'98, ...
Symbols	+, -, *, /, =, ...	
Others	URL, E-mail, ...	

Table 1 shows a brief summary of Not Standard Word (NSW)'s taxonomy based on one year People Daily Corpus. Since 95% of the NSW's in the corpus are number expression, including digit strings, digit strings combinations and symbols. Based on the combination of digit strings, this paper presents a NSW's taxonomy on five categories, as shown in Table 2.

Table 2.Categories of NSW

Single digit	100 天(day), 10%, ...
Double digit	100-200, 12:12, ...
Several digit	59.64.192.168 ...
Digit & English	Win98, p2p, ...
Symbols	@, \$, ...

The first category is composed of single direct digits with prefixes or suffixes like Chinese measuring units. Only one digit string can be found in this kind of NSW. The next category is for the NSW that has two digit strings connected together by a joint

symbol, such as "100-200". The third category should have three or more than three digit strings connected together. URL, Finance Stock could be found in this category. The fourth category is digit plus English words or letters. Chinese texts may contain English alphabets, which are typically read as a word or letter by letter in mandarin TTS system. However, English alphabets related with digit strings often have special meanings. The last category contains different kinds of symbols, such as "@, #". Some of them have various pronunciations. Based on the above five basic categories, in sum 50 types of NSW's are included in the taxonomy. Of all these NSW's, some have determined pronunciation, while some do not.

NSW's whose pronunciations are determined by formats are named as Basic NSW's (BNSW), and those with ambiguities are called Ambiguous NSW's (ANSW). BNSW have determined pronunciations while ANSW need further processing for the disambiguation. Table 3 shows some ANSW and their possible variation in pronunciation. Such NSW's need both internal and contextual information for disambiguation.

Table 3.Examples of ANSW

NSW	Say-as	Example
digits	digit by digit	2 米(metre)11
	integer	120
	pronunciation change	110
	english	p2p
slash	fraction	1/2
	not pronounce	T31/32
	date	2008/8
colon	time	时间(time)10:10
	rate	比分(score)10:10
year	some year	2008 年(year)
	many year	50 年(years)
sym-cross	multiple	身高(height)×体重(weight)
	substitution	周(first name)×

Finite State Automata (FSA) are designed to detect NSW's of all 50 above mentioned types and to give an initial classification based on NSW formats. Longer unit contains more information and thus has less ambiguity. Hence, Maximum Match strategy is adopted when nesting NSW's exist. That is, the longest NSW is considered as a NSW, not any of its substrings.

2.2. Rules and Maximum Entropy Classifiers

Maximum Entropy model presented in this paper contains six types, digits, year, hyphen, year-range, sym-hyphen and sym-slash. Rules cover most other types of ANSW.

The maximum entropy framework agrees with everything that is known, but carefully avoids anything that is unknown. In

other words, it estimates probabilities based on the principle of making as few assumptions as possible, given the imposed constraints. The probability distribution that satisfies the above property is the one with the highest entropy. It is of the following exponential form

$$P(y | x) = \frac{1}{Z(x)} \exp \left\{ \sum_i \lambda_i f_i(x, y) \right\}$$

Where x is a history or context, y is the outcome or category, and $Z(x)$ is a normalization function.

$$Z(x) = \sum_y \lambda_i f_i(x, y)$$

The features used in the maximum entropy framework are binary. Here's an example of a feature function, which implies the fact that digit string "110" has a pronunciation change.

$$f_i(x, y) = \begin{cases} 1 & \text{if } y = yb, nsw = 110 \\ 0 & \text{otherwise} \end{cases}$$

The training of maximum entropy model is to learn parameters λ_i . Parameter estimation methods include Generalized Iterative Scaling (GIS), Improved Iterative Scaling (IIS) and L-BFGS, etc. Data smoothing methods include Gaussian Prior and exponential prior.

The set of classifiers have both public and private features. Public features are shared by all classifiers while private features are designed for different classifiers. Public features are n -gram character features within size of 4:

Uni-gram: $Cn(n=-4,-3,-2,-1,0,1,2,3,4)$

Bi-gram: $CnCn+1(n=-4,-3,-2,-1,0,1,2,3)$

Tri-gram: $CnCn+1Cn+2(n=-4,-3,-2,-1,0,1,2)$

4-gram: $CnCn+1Cn+2Cn+3(n=-4,-3,-2,-1,0,1)$

Here, Cn can be a digit string, a character or a symbol.

Private features are different for each classifier. For example, the classifier of NSW type "yearange" contains two private features as follows:

if it has four figures

if it is more than three thousand years

Rules are designed for the ANSW except the above six types involved in the Maximum Entropy model. According to the FSA classifiers, we generate a sub-classification for each ANSW. Based on the FSA classification, such sub-classification usually has two types, regular pronunciation and special pronunciation. For example, "12:12" may be marked as type "colon" through FSA classifiers. Still, it is not easy to identify whether it should be read as "十二点十二分(*twelve twelve*)" or "十二比十二(*twelve to twelve*)". Thus, more complex rules depending on the neighboring context are needed for this disambiguation. In this paper, two sub-types of "colon" are discussed: "colon/tm" stands for timing pronunciation while "colon/rt" means that NSW should be read as ratio pronunciation.

2.3. Standard Words Generation

Standard word generation is the last module of text normalization which converts text to the corresponding Chinese-pronounced words. A set of systematic rules is applicable here.

The first step is to process some English words, which may appear in mandarin text. Once English words are met, a pair of brackets "(" will mark them. Some English words have special meaning as quantity units when their prefixes are digits. To convert these English units into Chinese pronunciation, a rule including 58 unit names is made. Other English words will be processed in English TTS.

The conversion module is a matching correspondence generation. In processing the tagged text, this paper offers a hierarchical way. As digits strings are the majority of all special words, "digits" type and "decimal" type make the basic level. Other levels can be treated as the arrangement and combination of these two types. The digits string consists of two parts: the integer part and the decimal part. The decimal part can be processed as reading digit by digit.

In Chinese, numbers are expressed in a way that every four digits are grouped and suffixed with the corresponding quantity. In contrast, number is counted three digits by three digits in English. For example, 123456789 in English is 123 million 456 thousand and 789, while in Chinese is 1 亿(*one hundred million*), 2345 万(*ten thousand*), 6789. In processing, first determine how many parts can the digits string divided by and process each part to generate standard pronunciation. The final step is to insert characters "万" and "亿" at the appropriate location.

The Following example is given to illustrate the processing method. In order to process the digits string "102034.567", find the comma and separate the string into "102034" and "567". "567" is read digit by digit "五六七(*five six seven*)". "102034" is separated into "10" and "2034". "10" and "2034" will be pronounced as integer "十(*ten*)" and "两千零三十四(*two thousand and thirty four*)", respectively. Finally "万" is inserted between the two parts to complete the reading as "十万两千零三十四(*one hundred and two thousand and thirty four*)".

3. EXPERIMENTS

Experiments are designed to test the performance of the text normalization. Experiments are implemented in BaiLing, a real TTS system authorized by FTR&D. Experiments here reflect the performance of whole text normalization process in TTS system.

3.1. Corpus

Experiments corpus is composed of 1000 sentences randomly extracted from People Daily Corpus. A dictionary trained by

FSA through regular expression and a ME model containing six types of NSW's are used in the experiments.

In the experiments, we use the SSML format in BaiLing TTS system. A "Say-as" element is used to indicate NSW's. The Chinese pronunciation and property of NSW's are also added in the "Say-as" element.

3.2. Experiments Results

Evaluation criteria are Precision and Recall. Precision shows the proportion between number of correctly tagged NSW and number of all tagged NSW while Recall indicates the ratio of correctly tagged NSW number and all manual tagged NSW number.

Baseline contains the methods of simple rules and traditional FSA+ME+Rules for mandarin text normalization. In table 4, the method of simple rules achieves the Precision of 90.77% and Recall of 88.49%. Moreover, the method of FSA+ME+Rules is 93.81% and 91.58 for Precision and Recall.

Table 4.Experiments Performance

	Precision	Recall
Simple Rules	90.77%	88.59%
FSA+ME+Rules	93.81%	91.58%
FSA+ME&Rules+Rules	96.02%	93.74%

As shown in Table 4, compared with the simple rules based method, the three-stage NSW's disambiguation strategy of FSA+ME&Rules+Rules improves the Precision and Recall by 5.25% and 5.15% respectively. The strategy also shows better performance than the three-stage method of FSA+ME+Rules: 2.21% higher of Precision and 2.16% higher of Recall.

It shows that the three-stage disambiguation strategy is much more effective than the simple rules based method for normalization. Moreover, compared with the traditional method for subclass disambiguation, the new classification combined with certain rules works well. The three-stage strategy is of great practicality in real TTS system.

3.3. Error Analysis

Some errors occur in the experiments. For example, "1.29" cannot be easily identified by the system, which means we don't know weather it stands for a date or a real decimal number. To mark up correctly, a possible complementary approach is the utility of neighboring tags or a more complex model, such as Condition Random Field (CRF).

4. CONCLUSION

This paper makes an extensive investigation of Chinese normalization. NSW's taxonomy is developed based on a Large-scale corpus. After a systematic analysis of the taxonomy,

a three-stage NSW's classification strategy is proposed, Finite State Automata for initial classification, Rules & Maximum Entropy classifiers for subclass disambiguation and General Rules for words conversion. Experiments results indicate that this approach is able to achieve a good performance with high precision and recall. Moreover, this approach of the three-stage NSW's disambiguation strategy for Chinese text normalization works well in real TTS system. It also shows great general application in further processing of TTS system.

REFERENCES

- [1] Richard Sproat, Alan Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, Christopher Richards, "Normalization of Non-Standard Words", *Computer Speech and Language*, 15(3):pp.287-333,2001
- [2] David Yarowsky, In Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg (eds.), "Homograph Disambiguation in Text-to-Speech Synthesis," *Progress in Speech Synthesis*, pp.159-175, 1996
- [3] Zhigang Chen, Guoping Hu, Xifa Wang, "Text Normalization in Chinese Text-to-Speech System," *Journal of Chinese Information Processing*, 17(4): pp.45-51, 2003
- [4] Andrew Breen, Barry Eggleton, Peter Dion, and Steve Minnis, "Refocusing on the Text Normalization Process in Text-to-Speech Systems," *In Proc. ICSLP 2002*, pp. 153-156, 2002
- [5] Xiaoru Wu, Renhua Wang, Guoping Hu, "Special text processing based external descriptor rule," Sixth International Conference on Spoken Language Processing (ICSLP2000) *In ICSLP-2000*, vol.1 pp.689-692, 2000
- [6] Xinnian Mao, Yuan Dong, Jinyu Han, Dezhi Huang, Haila Wang, "Inequality Maximum Entropy Classifier With Character Features For Polyphone Disambiguation In Mandarin TTS Systems", *IEEE International Conference on Acoustics, Speech, and Signal Processing* vol.4 IV-705-IV-708 ICASSP 2007
- [7] K. Panchapagesan, Partha Pratim Talukdar, N.Sridhar Krishna, Kalika Bali, and A.G.Ramakrishnan, "Hindi Text Normalization," *In Proc. KBCS 2004*, pp.19-22, 2004
- [8] Virongrong Tesprasit, Paisarn Charoenpornasawat and Virach Sortlertlamvanich, "A Context-Sensitive Homograph Disambiguation in Thai Text-to-Speech Synthesis," *In Proc. HLTNAACL 2003*, pp.103-105, 2003
- [9] M.H.Moattar, M.M.Homayounpour, and D.Zabihzadeh, "Persian Text Normalization Using Classification Tree and Support Vector Machine," *In Proc. ICTTA 2006*, pp.1308-1311, 2006
- [10] Yuxiang Jia, Dezhi Huang, Wu Liu, Yuan Dong, Shiwen Yu, Haila Wang, "Text Normalization in Mandarin Text-to-Speech System," *IEEE International Conference on Acoustics, Speech, and Signal Processing* pp.4693-4696, 2008